Mining Microdata: Economic Opportunity and Spatial Mobility in Britain, Canada and the United States, 1850-1911

Goals and objectives: For almost two centuries, social theorists have argued that differences in economic opportunity and geographic mobility on the two sides of the Atlantic led to fundamental differences in social structure. Nineteenth century commentators saw the United States and Canada as countries with great opportunity for economic mobility. In the twentieth century, Canadian and U.S. historians challenged this interpretation. Using linked censuses of more than a dozen communities, historians in both countries argued that despite high geographic mobility, nineteenth-century Canada and the United States had a rigid class structure with comparatively little upward mobility. But in the last decade the debate has turned again. Studies using national data have argued that the nineteenth-century United States was extremely fluid compared with nineteenth-century England. The new results suggest that there has been a dramatic decline in the United States in both economic and geographic mobility over the past 150 years. If confirmed, these results would have profound implications for our understanding of social structure and social change on both sides of the Atlantic.

We proposed to apply new data-mining technology to massive new census microdata collections in Britain, Canada, and the United States to address four key questions:

1. What were the relative levels of economic and geographic mobility in Britain, Canada, and the United States in the late nineteenth century?
2. What were the mobility trends in each country?
3. How were economic opportunity and geographic mobility interrelated in each country?
4. What individual and community characteristics were associated with economic and geographic mobility?

Challenges and lessons of international collaboration across disciplines and domains: We have found the most challenging aspect of collaboration to be the social one of coordinating the division of labour among three countries and five research sites. Our team draws on expertise in history, economics, sociology, demography, and computer science. We have used a combination of email, Skype, and file-sharing technology to work collaboratively between in-person meetings. The majority of the team have met twice in person at different institutions involved in the project, and smaller meetings of project members occur regularly. The lesson here is the obvious one that regular communication is important. Even in the digital age, in-person meetings fulfil a vitally important function of re-committing project partners to each other and the shared work.

A secondary, and related, challenge has been making decisions about data construction in a consistent way across countries. The underlying census data are similar, but not identical, meaning that we cannot create our data absolutely identically in all three countries. How does one make consistent outputs out of inconsistent inputs, and which trade-offs are acceptable? The record linkage of some population sub-groups may be hindered by their homogenous name pool, decreased literacy or failure to report age,

while the linkage of other sub-groups may be favoured by the greater birthplace granularity. These differences will be carefully documented in terms of potential biasing effects. This has been a creative tension that has sharpened our understanding of our methods and what we are trying to do. The lesson from this challenge reinforces the other lesson: regular communication in projects with multiple institutions is important.

Digital humanities and social sciences in big data projects: Our project has drawn on resources created originally by genealogical companies and societies: Family Search, findmypast.co.uk, and Ancestry. These organisations have donated the data to social scientists for research purposes. Through our work on Digging Into Data, and related projects, we have deepened our collaboration with these genealogical organisations on methods for identifying the same people consistently in multiple data sources. Though our precise objectives differ slightly from theirs—they are interested in maximizing potential links, we are interested in unbiased and representative links—the discussion across different research methodologies has been important for both groups.

Indicators of success: At the time of writing we have created 4 of the 6 datasets required for our analysis of changing social mobility. We are currently working on an analysis of social mobility in the first period (1850-1881), which will be applied to the second set of datasets (1880-1911) as soon as they are complete. Our production time for each subsequent dataset has declined as we have standardized methods across datasets.

Measuring impact: Our greatest impact to date has been through our collaboration with genealogical companies in working with the underlying datasets of complete count censuses. Our work with the data has helped identify problems in the census databases. Although our goals for linking people across censuses differ slightly as outlined above, we are sharing knowledge with the genealogical research community about ways to create linked data.

Knowledge dissemination mechanism and tools: Our datasets are being distributed at www.nappdata.org.

Importance of working with libraries, archives and data repositories: Our most important collaborations have been with genealogical societies and companies. The complete count datasets we are using for our research are expensive and time-consuming to create. By sharing resources and knowledge, we are able to reduce duplicated effort in building data resources that can be shared. The genealogical companies we are working with created the census databases for a private investment on which they want to earn a profit. Our collaboration with them is a good example of how academic and commercial organizations can find ways to partner on using data together. We have negotiated data usage agreements that make their data available to social scientists under conditions that safeguard the interests of both parties.

Capacity building: The principal investigators of the projects are historians and economists, with other faculty involved from sociology and demography. Our two post-docs on this project, Luiza Antonie (Guelph) and Tatiana Penkova (Leicester) have an information technology background. Their involvement in this project has increased their skills in applying information science methods to solve social science problems.